DOCUMENT RESUME

ED 399 291 TM 025 604

AUTHOR Hambleton, Ronald K.
TITLE Guidelines for Adapting Educational and Psychological Tests.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE Apr 96
NOTE 47p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Testing; *Psychological Testing; Scoring; *Test Construction; *Testing Problems; Test Use
IDENTIFIERS *International Test Commission

ABSTRACT

The International Test Commission formed a 13-person committee of psychologists representing a number of international organizations to prepare a set of guidelines for adapting educational and psychological tests. The committee has worked for 3 years to produce near final drafts of 22 guidelines organized into 4 categories: (1) context; (2) instrument development and adaptation; (3) administration; and (4) documentation and score interpretations. Each guideline by itself is described by a rationale for inclusion, a set of steps for achieving the guideline, a list of common errors, and references for follow-up study. This paper provides a report on the work of the committee in preparing the guidelines, the final version of which will be available in June of 1996. These guidelines must be incorporated into a set of ordered steps for conducting test adaptations. An abstract in French is attached. (Contains 4 figures and 27 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
RONALD K. HAMBLETON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 399 291

Guidelines for Adapting Educational and Psychological Tests[1,2]

Ronald K. Hambleton
University of Massachusetts at Amherst

Abstract

The International Test Commission formed a 13-person committee of psychologists representing a number of international organizations to prepare a set of guidelines for adapting educational and psychological tests. The committee has worked for three years to produce near final drafts of 22 guidelines organized into four categories: context, instrument development and adaptation, administration, and documentation/score interpretations. Each guideline itself is described by a rationale for inclusion, a set of steps for achieving the guideline, a list of common errors, and references for follow-up study. The purpose of this paper is to provide a report on the work of the committee in preparing guidelines. The final version of the test adaptation guidelines will be available in June of 1996.

---

[1]The main funding for this project has come from the National Center for Education Statistics of the Department of Education in the United States.

[2]Paper presented at the meeting of NCME, New York, 1996.

LR272

TM025604

BEST COPY AVAILABLE

Background

There is considerable evidence to suggest that interest in international comparative studies of educational achievement and cross-cultural research is growing. With this growth has come the need to adapt (commonly called "translate") achievement and aptitude tests and psychological instruments for use in multiple cultures and languages. For example, the International Association for the Evaluation of Educational Achievement (IEA) is conducting the third international mathematics and science studies (TIMSS) in over 40 countries. Over 30 different languages are represented in these 40+ participating countries. Major test adaptation efforts were carried out to insure that the assessments were equivalent in participating countries. In addition, for many years, popular American and British intelligence tests and personality instruments have been adapted for use into many languages and cultures. Professor Charles Spielberger (personal communication) reported, for example, that his state-trait measure of anxiety has been adapted for use in 45 languages. Substantially more adaptations might be expected in the future as (1) international exchanges of tests and instruments become more common, (2) credentialing exams are adapted for use in multiple languages (a likely consequence of the European Economic Community, for example), and (3) interest in cross-cultural research increases.

The technical literature for guiding the test adaptation process appears to be incomplete (from a measurement perspective), and scattered through a plethora of international

journals, reports, and books. Quite simply, there is no single complete source that practitioners can turn to for advice, nor has a complete set of guidelines for adapting tests ever been formalized. Also, the more complex measurement methods (e.g., item response models and structural equation models) which appear to be useful in formally establishing the equivalence of scores obtained from tests adapted for use in multiple languages and cultures do not appear to be well-known by researchers who do test adaptations (e.g., Hulin, 1987).

In view of the fact that "high-stakes" are often associated with the results from cross-cultural or international comparative studies of educational achievement, the need for professionally developed and validated guidelines for adapting tests and establishing score equivalence seems clear. For example, the results from recent international comparative studies of achievement have regularly been sighted by policy-makers in the United States as reasons for educational reform. Many other countries, based upon the high participation rate in TIMSS, probably share the United States interest in comparative achievement results. But these results, depend for their validity, on the suitability of the test adaptations that are made. A poor test adaptation can make a test more difficult or easier in a second language and can change the validity of the scores in significant ways.

Technical standards or guidelines for assessment practices concerning test development, reliability assessment, validity assessment and norming are available in many countries (see, for

example, the American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1985), but rarely has much attention been given to the preparation of guidelines for adapting tests and establishing score equivalence. For example, in the widely used AERA, APA, and NCME Test Standards only limited attention was given to the topic of test adaptations. And in Canada, which is a bilingual country, only three guidelines addressing test adaptations appear in the Canadian Psychological Association's test standards.

Purposes and Activities

The primary purposes of the International Test Commission (ITC)-initiated project were to prepare and to widely disseminate a set of guidelines for adapting tests and psychological instruments and establishing score equivalence across language and/or cultural groups. The term "adaptation" rather than "translation" was preferred by the test adaptation guidelines committee because the former term is broader and seemed to more accurately reflect the process of preparing a test or instrument for use in a second language or culture. Translation is always part of the adaptation process, but it is only one of several steps that must be carefully carried out to produce a test or instrument which is equally valid in two or more languages and cultures.

The work of developing test adaptation guidelines has been carried out by an international committee of psychologists working, to date, for three years. Activities completed include:

1. Obtained the cooperation of international groups with an interest in the project, and who could contribute technically to the process of developing guidelines. The list of participating organizations appears in Figure 1.
2. Obtained funds for logistic support of the project and travel support for two committee meetings. Logistic support funds are coming from the National Center for Education Statistics (NCES), of the U.S. Department of Education. Travel funds have come from many sources, including NCES, and all of the participating organizations.
3. Prepared a background paper on test adaptations and a plan for completing the project (see, Hambleton, 1993).
4. Implemented the plan with the assistance of a 13-person committee representing major international organizations.
5. Presented drafts of the guidelines at the annual meeting of the American Psychological Association in Toronto in 1993 and the 23rd International Congress of Psychology in Madrid in 1994 and revised the guidelines based upon the feedback received (Hambleton, 1994; van de Vijver & Hambleton, in press).

Insert Figure 1 about here.

The thirteen committee members met at the IEA Headquarters in the Hague in September of 1992 and (1) agreed on the scope of the project, (2) discussed the main issues in adapting tests and

psychological instruments, and (3) formed work teams to complete the project. An outline of the final report and a very preliminary version of the guidelines were prepared. Following the meeting, the work continued in drafting the guidelines and the final report and the supporting documentation for the guidelines. A second meeting was held in the spring of 1994, again at the IEA Headquarters in the Hague, to review drafts of the guidelines and to edit a draft of the final report. At this stage of our work, a draft of the guidelines and the final report are available, and the reviews of 40 psychologists have been collected and are being used to finalize the guidelines.

The last activity, which will be completed later this year and next, is

6. Disseminate the guidelines in a number of ways, including (a) a final report, (b) international conference presentations (for example, a symposium has been organized for the International Congress of Psychology in Montreal in 1996), (c) several journal articles (see, for example, Hambleton, 1994; van de Vijver & Hambleton, in press), (d) summaries of the guidelines in various international journals and newsletters, and (e) an international conference in 1997 (Professor Thomas Oakland from the University of Texas in the United States and Vice-President of the International Test Commission is the co-ordinator) on the issues and methods associated with test adaptations.

The remainder of this paper is divided into three sections: First, the members of the committee will be introduced, all of

whom contributed to the contents of this paper. Second, and most importantly, the current guidelines and rationales for their inclusion will be presented. In a final section, some preliminary conclusions about the work and future initiatives will be presented.

Committee Composition

The International Test Commission (ITC) is coordinating the project and nominated four committee members. The chairperson of the committee is Ronald K. Hambleton from the University of Massachusetts at Amherst in the United States and, since the summer of 1994, the Past-President of the ITC. The European Association of Psychological Assessment (EAPA), the International Association of Applied Psychology (IAAP), the International Association for Cross-Cultural Psychology (IACCP), IEA, the International Language Testing Association (ILTA), and the International Union of Psychological Science (IUPsyS) provided nominations for committee membership.

In order that the committee not be dominated by members from one country, or from a narrow technical perspective, several of the participating organizations were asked to provide the names of four or five possible representatives from which the committee selections could be made. ITC's representatives were to be chosen in much the same way. Persons were selected because of special expertise and experiences they could bring to the committee. For example, it was desirable to have a member of the committee who had participated in a project to develop technical standards for tests and testing practice. A person with

experience in personality measures was valuable, as was a person who had used item response theory models and other statistical methodologies successfully in test and psychological instrument adaptation work. To round out the committee, the European Test Publishers Group (ETPG) was invited to serve on the committee and provide a test publishers' perspective.

Names of members on the committee and the organizations they represent appear in Figure 2. Representatives from the Far East and South America were not included on the committee because of the costs involved in bringing these representatives to the two meetings. We did, however, invite psychologists from the Far East and South America to serve as reviewers of drafts of the guidelines. At the present time, 40 psychologists have participated as reviewers and several were from the Far East and South America.

Insert Figure 2 about here.

Instrument Adaptation Guidelines

The guidelines for adapting educational and psychological tests or instruments were organized into four sections: context, instrument development and adaptation, administration, and documentation/score interpretations. (We will use the terms "tests" and "instruments" interchangeably throughout the paper.) Our thinking was that the guidelines would be more convenient to use if they were organized into some meaningful categories. The committee felt that four categories made good practical sense.

Guidelines in the context category address concerns about construct equivalence in the language groups of interest. In fact, adaptations are of no value if construct equivalence cannot be established. The instrument development and adaptation category includes the guidelines which arise in the process of adapting an instrument, everything from choosing translators to statistical methods for analyzing empirical data to investigate score equivalence. The third category, administration, addresses guidelines having to do with the ways in which instruments are administered in multiple language groups, and this includes everything from selecting administrators, to the choice of item formats, to establishing time limits. The fourth category of guidelines concerns documentation and score interpretations. Typically, researchers provide very little documentation of the adaptation process to establish the validity of an adapted test, and misinterpretations of scores from instruments in multiple languages are common. The guidelines address concerns in these areas.

The following definition of an instrument adaptation guideline was adopted by the committee:

> An instrument adaptation guideline is a practice which is judged as important for conducting and evaluating the adaptation or parallel development of psychological and educational instruments for use in different populations.

The 22 guidelines advanced by the instrument adaptation guidelines committee are summarized below and also in Figure 3,

along with a rationale for the inclusion of each guideline. In the final report, each guideline is described by not only a rationale for including the guideline, but also steps for addressing the guideline in practice, a list of common errors, and a set of references. A complete example of one of the guidelines, D.1, is given in Figure 4.

Insert Figures 3 and 4 about here.

Context

C.1 Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.

Rationale/Explanation. There are many factors which affect cross-cultural/language comparisons that need to be considered whenever two or more groups from different language/cultural backgrounds are compared, especially when an instrument is being developed or adapted, or scores are being interpreted. However, often it is necessary that some of these factors are not merely taken into account, but that practical steps be taken to either minimize or eliminate the likely (unwanted) effects of these factors on any cross-cultural/language comparisons that are made.

For example, the different levels of test motivation of participants from Korea and the United States in a recent International Assessment of Educational Progress study can be regarded as one of the probable reasons for the vastly different performances of participants from these two countries (Wainer,

1993). The Koreans, who performed extremely well, were made aware of the great honor of being chosen (randomly) to represent their school and country and thus had a greater responsibility (and motivation) to perform well compared to the Americans, who were selected to participate in "just another study." To at least some American students, the assessment was an inconvenience.

There is another point to be made. Lonner (1990) noted that in the developed and "highly psychologized" nations, the typical citizen is familiar with many test taking practices; i.e., the desirability of optimum performances on ability tests, or honest responses on psychological tests. However, for many persons in developing nations, since testing is <u>not</u> part of the cultural landscape, expected test behaviours should not be assumed. Steps must be taken to insure that all participants are working under the same set of assumptions and expectations about the assessment.

C.2 The amount of overlap in the constructs in the populations of interest should be assessed.

<u>Rationale/Explanation</u>. Differences that exist between various cultural and language groups are a function of not only the different traditions, norms and values, but of different world views and interpretations as well. It is thus entirely possible that the same construct is interpreted and understood in completely different ways by two different groups. For example, the concept of 'intelligence' is known to exist in almost all cultures. However, in many Western cultures this concept is

associated with being able to produce responses very quickly, while for many Eastern cultures, intelligence is often associated with slow thoughtfulness, reflection, and saying the right thing (Lonner, 1990).

Cross-cultural researchers, especially, have to ensure that the constructs measured by an instrument in the original source cultural/language group can be found in the same form and frequency in the other groups under investigation. In fact, Butcher and Garcia (1978) noted that it is important to assess whether the construct itself is even meaningful in the target group.

Van de Vijver and Poortinga (1991) noted the studies by Greenfield (1966, 1979) as an example where the construct being measured was familiar in another cultural group, but had an added meaning as well. Based on his (western) understanding of the word "more", the study conducted by Greenfield (1966, 1979) concluded that the concept of the principle of conservation was not mastered by the Wolof people. However, what he (Greenfield) was not aware of was that in the Wolof language, the word "more" referred to both quantity and level, and thus the conclusion drawn by the researchers was completely misleading. Such a situation could have easily been avoided had the researcher first ascertained the meaning of the relevant concept as understood by the group under investigation.

Instrument Development and Adaptation

D.1 Instrument developers/publishers should insure that the adaptation process takes full account of linguistic and

cultural differences among the populations for whom adapted versions of the instrument are intended.

Rationale/Explanation. The rationale for this guideline along with the other parts of this guideline description are contained in Figure 4.

D.2 Instrument developers/publishers should provide evidence that the language use in the directions, rubrics, and items themselves as well as in the handbook are appropriate for all cultural and language populations for whom the instrument is intended.

Rationale/Explanation. One of the causes of poor instrument development for cross-cultural research is that the source language version is often unnecessarily complicated and therefore quite difficult to translate accurately. Another cause may be that concepts, expressions, and ideas used in the source language version of the instrument do not have equivalents in the target language.

Also it is important to ensure that the vocabulary used for instruments in two or more languages is comparable in terms of the level of difficulty of words, readability, grammar usage, writing style and punctuation. In this context, the reasons for using the instrument, for example, assessment of adult literacy, and the reading level of participants (children versus adults) should be carefully considered.

An additional factor to consider in this context is that exposure to, or having "partial knowledge" of the source language can make the interpretation of instrument questions easier or

harder for some groups within a population. This situation can easily occur in countries with population groups whose home (mother) language is different from that of the "official" or dominant language.

D.3 Instrument developers/publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and procedures are familiar to all intended populations.

Rationale/Explanation. Specific formats (e.g. multiple choice, essay) and certain conventions and procedures in giving instructions and presenting test items may not be equally familiar to all populations. Conventions and procedures range from language use in test rubrics, lay-out and use of graphics, and presentation mode (e.g., paper & pencil, computer). To ensure fairness it is important that all formats, conventions, and procedures be familiar to all populations for whom adaptations of the instrument are intended and this may involve extensive practice materials to reduce bias due to unfamiliarity of some aspect of the assessment process. Another option might include the use of different assessment approaches in different groups.

D.4 Instrument developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.

Rationale/Explanation. Any adapted instrument that proves easier or more difficult to read or understand because of the specific content may introduce an additional source of bias.

Larson (1987) noted that the culture of the source text determines the meaning of the text while the culture of the target language speakers affects how they understand the translation.

In various parts of the world different units are used to express quantity in, for example, weight, length, and money. An adaptation of a test can be become more difficult for the target population if the units used are less familiar or if they require different mathematical operations.

Also, certain stimulus material (diagrams, tables, figures, famous landmarks) may not be equally familiar to all populations.

D.5 Instrument developers/publishers should implement systematic judgmental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and compile evidence on the equivalence of all language versions.

Rationale/Explanation. Care is needed to ensure the equivalence of meaning in questions/tasks/rating scales in different languages and cultures. Judgmental methods of establishing translation equivalence are based on a decision by an individual or population of individuals on the degree of each item's translation equivalence. The two most popular designs are forward translations and backward translations.

With forward translations, a single (or a group of) translator(s) first translates or adapts the instrument from the source to the target language. Equivalence of the two versions is then judged by another group of translators, and any changes or revisions can be made, if needed. In the back translation

design, an instrument is first translated or adapted into the target language, then re-translated or adapted from the target into the source language by a different group of translators. A judge or group of judges assess the equivalence of an instrument by comparing the two source language versions. In practice, judgmental methods should be used as preliminary checks of translation equivalence before an instrument is administered and statistical methods are applied.

D.6 Instrument developers/publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish item equivalence between the different language versions of the instrument.

Rationale/Explanation. The data collection design refers to the overall method by which equivalence between different adapted instruments is assessed. A first requirement with respect to the data collection is that samples should be sufficiently large to allow for meaningful interpretation. Though this requirement holds for any type of research, it is particularly relevant in the context of instrument adaptation, because the statistical techniques needed to establish equivalence and to reject hypotheses on item or test bias can only be applied meaningfully with sufficiently large samples.

In many cross-cultural and especially cross-ethnic studies, the sample sizes for the target populations tend to be very much smaller than the source population. Often it is not possible to conduct reliable and valid analysis on populations with small sample sizes as most statistical techniques require large

samples. This situation can have a direct influence on the types of statistical analysis that can be carried out. Structural equation modelling, item response theory, item bias studies, etc., are easier to carry out with large samples in both the source and target language/cultural populations.

This is especially true in the case of item bias studies, where "minority populations" typically consists of small samples (Hambleton, et al, 1993; Linn & Harnisch, 1981). On of the consequences of this is that these "minority populations" are generally not included in the analysis as items on test instruments cannot be reliably pretested to detect any non-equivalence and/or bias. Another consequence of small sample sizes in target populations is that the power of statistical techniques to detect bias is greatly reduced and thus many bias items are missed (Swaminathan & Rogers, 1990).

The design for the empirical study should be a function of variations in (1) the nature of the participants (monolinguals or bilinguals), (2) the version of the instrument (original, adapted or back-adapted) used, and (3) the specific statistical technique applied (discussed in greater detail in Guideline D.7). Three data collection designs used to establish item equivalence of an instrument in different languages are common:

a. Bilingual examinees take source and target versions of the test. In this design, the same participants take both the source and target versions of the instrument. The advantage of this design is that differences in participant characteristics on the instrument (e.g., ability differences) can be controlled.

However, the design is based on the assumption that participants are bilinguals who are equally proficient in each of the languages. This is highly unlikely to occur for a substantial number of individuals (Cziko, 1987; Rosansky, 1979) and the assumption requires to be tested thoroughly prior to the experiment proper.

A second major problem with this approach is that results may not be generalizable to the intended populations as bilingual participants tend to be, on the average, more capable that their monolingual counterparts (Hambleton, 1993). This design is often best implemented with another design so that convergent validity of results can be addressed.

A variation of this bilingual design, which has the same limitations, but which is easier to implement, involves randomly assigning bilingual participants to take one of the versions of the instrument. In this case, a randomly-equivalent populations design is in effect.

b. Source language monolinguals take the original and back-translated versions. This design involves the administration of the original and back-adapted versions of the instrument to a sample of monolingual participants in the source language. Item equivalence is identified by comparing participant performances on the different versions of the same item. The advantage of this design is that by using one sample of participants, the resulting scores are not confounded by differences in participant characteristics (Hambleton & Bollwark, 1991).

Two major shortcomings, however, threaten this design. First, no empirical data are collected on the target language version of the instrument. That is, no target language monolinguals are used even though the aim is to generalize the meaning of scores to the population of target language monolinguals. Secondly, the achievements on the subsequential source language versions may not be independent, because it cannot be ruled out that learning resulting from administering the first original source language version influences the results on the back-translated version. Counter-balancing can reduce the significance of practice effects.

c. Source language monolinguals take source language and target language monolinguals take target language. In this design, source and target language monolinguals are used, with each taking the version of the instrument in their respective languages. The main advantage of this approach is that results obtained are more generalizable to the respective populations (Hambleton & Bollwark, 1991). The main problem with this method is that since two different populations of participants are used, the resulting scores may be confounded with differences in participant characteristics between the two samples. However, this problem can be controlled for by matching samples or using statistical techniques that condition on characteristics measured by the instrument (e.g., ability) when comparing examinees.

D.7 Instrument developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the instrument, and (2) identify

problematic components or aspects of the instrument which may be inadequate to one or more of the intended populations.

Rationale/Explanation. Statistical techniques provide useful information for assessing the equivalence of instruments developed in more than one language (Hambleton & Kanjee, 1995). These techniques should be used to supplement judgmental techniques as they are able to identify non-equivalent instruments that may not be readily detected when using judgmental designs. Another advantage is that statistical techniques elicit information directly from the participants, within the context of an actual instrument administration, and are thus extremely useful for identifying scales that might pose problems in practice.

The decision to use any specific statistical technique depends on whether (1) a common scale is assumed for the various versions of the instrument, and (2) procedures used are conditional or unconditional (van de Vijver & Poortinga, 1991). If a common scale is assumed, the results from the different versions of the instrument are directly comparable, as test scores have a similar meaning. In some situations, however, it may be necessary to determine, rather than just assuming, the existence of a common scale. In these situations, the use of factor analytic methods can prove to be quite useful (Poortinga, 1983).

When conditional procedures are used, equivalence is assessed based on the level of participant characteristics

measured by the instrument. It is assumed that non-equivalence affect participants at particular ability levels differently as non-equivalence may not be invariant across the whole range of scores. Thus, in this technique, participants are divided according to the level of their raw scores and equivalence is analyzed per level. For example, participants at the lower level of ability may respond differently from participants at the higher level of ability. Unconditional procedures, on the other hand, simply entail a direct comparison of the statistics between populations compared. It needs to be noted that both these procedures are based on the assumption that a common scale exists between the populations compared. Of these two procedures, conditional procedures are preferred as they yield more detailed information (Hambleton & Kanjee, 1995; Hambleton, et al., 1993; van de Vijver & Poortinga, 1991).

D.8 Instrument developers/publishers should provide information on the evaluation of validity in all target populations for whom the adapted versions are intended.

Rationale/Explanation. Existing instruments are often developed and standardized for use in one culture and adapted for use in another culture. The advantage is that time and expense can be saved if existing instruments are adapted (Hambleton & Bollwark, 1991; Brislin, 1986). In addition, (1) there is usually data available against which comparisons can be made, and (2) there is an added sense of security when well established instruments are used (Brislin, 1986). However, many constructs that are taken for granted in the Western world, and for which

many instruments have been specifically developed, may not exist at all in other cultures. If they do exist, their behavioral manifestations and interpretations may vary considerably (Lonner, 1990). Construct validity evidence must be compiled in each population where the instrument will be used. As is well-known, a construct validity investigation is time-consuming to plan and carry out because it is typically extensive, and involves a variety of studies and methodologies including intertest, intratest, criterion-related, experimental, and multi-trait multi-method.

One disadvantage of adapting existing instruments for use in another culture is that it is possible for certain aspects of a phenomenon as defined and used by people in other language/ cultural populations to be missed (Brislin, 1986). Another disadvantage, is that if the constructs measured in the 'original' population are <u>not</u> found in the target culture, or are defined and used in some different form and frequency, the resulting scores can prove to be extremely misleading. Thus it is crucial to first determine the validity of the construct being assessed in the target language before an instrument is adapted.

D.9 Instrument developers/publishers should provide statistical evidence of the equivalence of questions for all intended populations.

<u>Rationale/Explanation</u>. One of the most important statistical analyses in validating an instrument for use in two or more cultural or language populations is an item bias study or referred to currently as a <u>differential item functioning</u> (DIF)

study. Basically, support for the equivalence of an instrument for two populations requires that there be evidence that when members of the two populations have equal ability, the two should perform in an equivalent fashion on each item. When performance is not equivalent, a sound reason must be available or the item should be deleted from the instrument. This does not mean that there cannot be overall performance differences on the instrument. In general, differences should be expected. What it does mean is that when members of the two populations are matched on the construct measured by the instrument, if differences exist, then DIF is present and the properties of the item must be studied carefully prior to any use of the item in the instrument.

This type of study involves a careful look at the item performance for members of the two populations of interest (e.g., citizens of China and South Africa) matched on the ability measured by the instrument. This means that instead of comparing (say) item difficulty values of samples of Chinese and South Africans, comparisons in performance are made between Chinese and South Africans samples of equal ability. The main idea is that if members of the two populations are of equal ability, then their performance on a task or test item should be equal except for any sampling errors due to sample size. When population differences in performance beyond sampling errors are noted, the item is labelled DIF and more intensive investigations are carried out to identify the reason for the differences. Differences in population performance are investigated for various levels of ability along the ability scale. Items flagged

as "DIF" may be problematic because of a poor translation or because of the use of a term, situation, or expression that is unknown or unfamiliar to one of the populations. Many other possibilities exist, too. Perhaps the skill measured by the item is not part of the repertoire of the target language population, or perhaps the item format is unfamiliar. Determining the reason for the difference is important because it influences the ultimate determination of what to do with the item.

This guideline can be meaningfully addressed once there is evidence that the construct is relevant in the populations of interest, and there is evidence that the translations or adaptations have been carefully checked (perhaps through a back-translation design). Basically, there are three methodologies that can be used to conduct the types of analyses required by this guideline: (1) item response theory procedures (see, for example, Ellis, 1989), (2) Mantel-Haenszel (MH) procedure and extensions (see, for example, Hambleton, et al., 1993; Holland & Thayer, 1988; Holland & Wainer, 1993), and (3) logistic regression (LR) procedures (Swaminathan & Rogers, 1990). All of these methodologies are "conditional" in the sense that comparisons are made between groups of persons (e.g., English and French) who are assumed to be "matched" on the ability or abilities measured by the instrument. With IRT procedures, examinees are matched using estimated ability scores (estimated using the item score patterns). With the other two procedures, the total instrument score (or a score adjusted by deleting questionable items) is used to match examinees. All three

methodologies can produce reliable and valid results providing the sample sizes are of substantial size and they are implemented correctly and the results are interpreted carefully. Sample sizes of about 200 per population are needed for the MH and LR procedures. In general, substantially larger samples are needed with the IRT procedures (though the Rasch model requires sample sizes equivalent to the other two procedures).

D.10 Non-equivalent questions between versions intended for different populations should not be used in preparing a common scale or in comparing these populations. However, they may be useful in enhancing content validity of scores reported for each population separately.

Rationale/Explanation. Questions on adapted instruments are identified as non-equivalent either because they are poorly adapted or culturally inappropriate (Hulin, 1987). These questions cannot be used as a basis for comparisons as they provide different information for the populations being compared. Poorly adapted questions can either be revised (if the intention is to use them again) or eliminated. However, well-adapted questions that are identified as non-equivalent (or culturally inappropriate), can often provide useful additional information about the specific populations being compared. Identifying the source of non-equivalence of these questions can provide further insight about the respective cultural/language populations that could increase our understanding of that population (Ellis, 1991).

Administration

A.1 Instrument developers and administrators should try to anticipate the types of problems that can be expected, and take appropriate actions to remedy these problems through the preparation of appropriate materials and instructions.

Rationale/Explanation. In general, the number of administration problems to be expected will vary as a function of the cultural and linguistic distance between the population groups involved or between the culture in which the instrument was first applied and the culture in which the instrument will be applied. Knowledge of the culture and language of the target groups is required. The developer is expected to address explicitly the problems most likely to affect comparability and to discuss actions that should or might be taken. Empirical evidence should be presented to support a claim of comparability. If this is not possible, justification should be put forward for the cross-linguistic use of the instrument.

A.2 Instrument administrators should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores.

Rationale/Explanation. Experienced instrument developers have a firm background in intra-cultural test development. However, additional experience is required to become sensitive to the intricacies and peculiarities of cross-cultural instrument administration. Specific factors that require attention in the administration of an instrument for a particular group are

difficult to define in general terms. A practical approach is to provide a list of frequently occurring problems and validity threatening factors. For instrument administration, a thorough knowledge of the linguistic and cultural aspects of the target group is indispensable. For example, three or four points on a rating scale in Turkey seems to be optimal (according to some psychologists in that country who reviewed our guidelines). With more points, semantics become problematic.

A.3 Those aspects of the environment that influence the administration of an instrument should be made as similar as possible across populations for whom the instrument is intended.

Rationale/Explanation. Instrument administration conditions can be a source of unintended score variation. In order to maximize the validity and comparability of test scores across cultural groups, possible sources of score variation should be described.

A.4 Instrument administration instructions should be in the source and target languages to minimize the influence of unwanted sources of variation across populations.

Rationale/Explanation. Cross-linguistic research will often deal with groups with a dissimilar cultural background. When the subjects or clients begin to answer the actual test questions/tasks/ratings the influence of unwanted sources of intergroup differences (disturbing factors) should be reduced as much as possible. The instructions serve this purpose. Good

instructions anticipate and attempt to reduce the effect of the disturbances.

A.5 The instrument manual should specify all aspects of the instrument and its administration that require scrutiny in the application of the instrument in a new cultural context.

Rationale/Explanation. Many aspects that are presumably relevant to the application of an instrument in other linguistic groups can be anticipated by the instrument developer: During the development and the validation of an instrument, developers have gathered valuable information about the specific issues that could be relevant in instrument adaptations and the application of the instrument in other linguistic groups. In some cases, the developer will even have data obtained among cultural minorities or cross-cultural applications available. Relevant information on the administration in these cultural groups should be provided.

A.6 The administrator should be unobtrusive and the administrator-examinee interaction should be minimized. Explicit rules that are described in the manual for the instrument should be followed.

Rationale/Explanation. The influence of the administrator on the measurement outcome can be substantial. As far as the behavior of the administrator is aimed at enhancing the standardization of the administration, the influence is desirable. However, the administrator can also have a less obvious and undesirable influence. In addition, administrator characteristics such as gender, age, or even style of clothing

can influence the measurement outcome. If a newly developed or adapted instrument is applied in a cultural group it will be relatively easy, possibly with the help of local informants, to pinpoint administrator characteristics that might endanger the validity of the instrument outcome. Appropriate actions (such as a small pilot study) can then be taken. Particularly in the case of a dissimilar cultural background between administrator and examinee, the potential obtrusiveness of the administrator should be scrutinized. It is the task of the instrument developer to design instruments in a way that the possible effects of the administrator's person is minimized. Furthermore, various administrators' characteristics such as gender, ethnic background, or use of a particular dialect of language can effect the examinee's behavior. It is important for the administrator to establish a good working relationship with the examinee but regulating this relationship should not challenge the measurement process.

Documentation/Score Interpretations

I.1 When an instrument is adapted for use in another population, documentation of the changes should be provided, along with evidence of the equivalence.

Rationale/Explanation. For many measurement specialists, as well as users of (adapted) instruments, information regarding the adaptation of an instrument can provide a great deal more insight about the suitability of using the instrument within a specific context. For example, knowledge that certain cultural (economical, social, etc.) factors were taken into consideration

in the adaptation of an instrument for Spanish speakers in a South American country can be extremely useful in deciding whether the instrument could be of similar use for Spanish speakers in the U.S. In this context, the entire procedure followed to adapt the instrument should be fully documented in the manual so as to facilitate evaluation of the instrument by other users. The documentation should include a detailed, step by step account of the entire procedure, including the design used, methods employed to assess equivalence between the adapted or translated versions, identification, selection and use of translators, the reasons and justifications for the use and inclusion of items as well as information about those items that were modified or not included, some of the major problems encountered and how they were solved, all aspects relating to the administration of tests including the selection and training of administrators, and the interpretation of results. This guideline is one of the most important.

I.2 Score differences among samples of populations administered the instrument should <u>not</u> be taken at face value. The researcher has the responsibility to substantiate the differences with other empirical evidence.

<u>Rationale/Explanation</u>. The common error is to be rather casual about the test adaptation process, and then interpret the score differences among samples of populations as if they were real. This mindless disregard of test adaptation problems and the need to validate instruments in the cultures where they are used has seriously undermined the results from many cross-

cultural studies. A technically sound test adaptation project is valuable in contributing to the validity of the adapted instrument. On the other hand, researchers must still make every effort to interpret their findings from multiple versions of an instrument carefully. This means, for example, that corroborating evidence should be compiled whenever possible, and when it can't be, extreme caution should be shown in interpreting results obtained in different populations.

I.3 Comparisons across populations can only be made at the level of invariance that has been established for the scale on which scores are reported.

Rationale/Explanation. Sometimes it is possible to place the scores from different language versions of a test or instrument onto a common scale. With access to large samples, and powerful statistical models such as those from item response theory (see, for example, Hambleton, Swaminathan, & Rogers, 1991), complex "equating" of scores from adapted versions of a test is possible when the construct is "reasonably equivalent" across the multiple versions of the test and the appropriate data are available (for example, see D.6). When this is possible, all types of comparisons of scores can be made including means, standard deviations, and distributions. But often scores from different language versions of a test have not been properly equated and then scores cannot be directly compared. Still, comparisons can be made about the role of the construct in each language version. For example, for an aptitude test adapted from English to French, a researcher may be interested in comparing

the predictive validity of the test in each language group. The main point of this guideline is to insure that researchers do not make unwarranted comparisons of scores from multiple versions of a test, and that they limit their interpretations to those for which validity evidence is available.

I.4 The instrument developer should provide specific information on the ways in which the socio-cultural and ecological contexts of the populations might affect performance on the instrument, and should suggest procedures to account for these effects in the interpretation of results.

Rationale/Explanation. In any cross-national/cultural study, the different factors that are relevant to the purpose for assessment need to be considered to gain a complete understanding of the results (Bracken & Barona, 1991). The different socio-political factors that invariably affect performance on the instrument are too often taken for granted (van de Vijver & Poortinga, 1991). For example, when comparing academic performance of students from developing and developed nations, or mainly industrialized and mainly rural societies, differences in performance may not be related to lack of ability but rather to a lack of access to resources, or it may be a reflection of the quality of educational services available. Other factors that could prove relevant include educational policies, expenditure on education, curricula, access to schooling, class sizes, availability of proper equipment and facilities, home language vs. language of instruction, teacher qualifications, political climate of assessment, literacy rate, etc. However, many of

these factors can quite easily be considered as external, [seemingly unrelated] factors, and are often only known to those very familiar with the culture/nationality. This makes it more difficult to identify these factors and thus emphasizes the need for translators well versed in not only the language but the culture as well.

Conclusions

The guidelines offered above are not final. Minor changes are still expected. Reviews of earlier drafts continue to arrive. Our hope is that the guidelines and associated descriptions will be useful to the many organizations participating in the test adaptation process and contribute to the validity of cross-language and cross-cultural research. In an earlier section, dissemination efforts were highlighted. We also anticipate some new ventures.

One venture is to compile a set of validated steps for practitioners to go through in adapting tests and establishing test score equivalence. In other words, the 22 guidelines need to be integrated into a set of ordered steps for conducting test adaptations. These steps might include such activities as (1) determining the need for a test adaptation project, and establishing the likely equivalence of the construct of interest in the multiple languages and/or cultures, (2) selecting translators and choosing a translation design, (3) implementing the design and making appropriate revisions to the adapted instrument, (4) determining the expected uses of the instrument and then designing studies to compile validity evidence (e.g.

item bias studies, factor analytic studies, item analysis, criterion-related studies), (5) if a common scale is needed, designing, and carrying out a study to place scores on a common scale, and (6) documentation of the full process and the validity evidence. In future work, these steps will be delineated, and examples of successful test adaptation projects will be compiled.

Author's Address:

Ronald K. Hambleton
School of Education
152 Hills South
University of Massachusetts
Amherst, MA 01003-4140
U.S.A.
Telephone: (413) 545-0262
FAX: (413) 545-4181
E-mail: RKH@EDUC.UMASS.EDU

Résumé

La Commission Internationale des Tests a constitué un comité de 13 psychologues, représentant diverses organisations internationales, dans le but de définir un ensemble de principes pour l'adaptation de tests éducatifs et psychologiques. Aprés deux années de travail, ce comité a élaboré un première version d'un document présentant 22 principes organisés en quatre catégories: contexte, développement et adaptation d'instrument, administration, et documentation/interprétation des scores. Chaque principe est lui-même décrit par: une justification de son inclusion dans la liste des principes, un ensemble d'étapes pur atteindre ce principe, une liste d'erreurs fréquentes, et des références pour un étude complémentaire. L'objectif de cet article est de présenter l'état d'avancement du travail du comité. La version finale des principes sera disponible dans le courant de l'année 1995.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. School Psychology International, 12, 119-132.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), Field methods in cross-cultural psychology (pp. 137-164). Newbury Park, CA: Sage Publications.

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. The Personnel and Guidance Journal, 56(8), 472-475.

Cziko, G. (1987). Review of the Bilingual Syntax Measure I. In J. C. Alderson & K. J. Krahnke (Eds.), Reviews of English Language Proficiency Tests. Washington, DC: TESOL.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. Journal of Applied Psychology, 74(6), 912-921.

Ellis, B. B. (1991). Item response theory: A tool for assessing the equivalence of translated tests. Bulletin of the International Test Commission, 18, 33-51.

Greenfield, P. M. (1966). On culture and conservation. In J. S. Bruner, R. R. Oliver, & P. M. Greenfield (Eds.), Studies in cognitive growth (pp. 225-256). New York: Wiley.

Greenfield, P. M. (1979). Response to Wolog "magical thinking." Journal of Cross-Cultural Psychology, 10, 251-256.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229-240.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. Bulletin of the International Test Commission, 18, 3-32.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. European Journal of Psychological Assessment, 9(1), 1-18.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. European Journal of Psychological Assessment, 11, 147-160.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (Eds). (1993). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Publishers.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology, 18, 115-142.

Larson, M. L. (1987). Establishing project specific criteria for acceptability in translations. In M. G. Rose (Ed.), American Translators Association Scholarly Monograph Series: Vol. 1: Translation excellence: Assessment, achievement, maintenance (pp. 69-76). Binghampton, NY: University Center at Binghampton (SUNY).

Linn, R. L., & Harnisch, D. L. (1981). Interaction between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Lonner, W. J. (1990). An overview of cross-cultural testing and assessment. In R. W. Brislin (Ed.), Applied cross-cultural psychology (pp. 56-76). Newbury Park, CA: Sage Publications.

Poortinga, Y. H. (1983). Psychometric approaches to intergroup comparison: The problem of equivalence. In S. H. Irvine & J. W. Berry (Eds.), Human assessment and cross-cultural factors (pp. 237-258). New York: Plenum Press.

Rosansky, E. J. (1979). A review of the Bilingual Syntax Measure. In B. Spolsky (Ed.), Some major tests: Advances in language testing. Series 1. Arlington, VA: Center for Applied Linguistics.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Van de Vijver, F. J. R., & Hambleton, R. K. (in press). Translating tests: Some practical guidelines. European Psychologist.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), Advances in educational and psychological testing (pp. 277-307). Dordrecht: Kluwer Academic Publishers.

Wainer, H. (1993). Measurement problems. Journal of Educational Measurement, 30, 1-21.

International Test Commission (ITC)

European Association of Psychological Assessment (EAPA)

European Test Publishers Group (ETPG)

International Association for Cross-Cultural Psychology (IACCP)

International Association of Applied Psychology (IAAP)

International Association for the Evaluation of Educational Achievement (IEA)

International Language Testing Association (ILTA)

International Union of Psychological Science (IUPsyS)

Figure 1. Participating organizations in the development of instrument adaptation guidelines.

Chairperson

Ronald K. Hambleton (ITC)
University of Massachusetts at Amherst, U.S.A.

Committee Members

Glen Budgell (ITC)
Canadian Nurses Association, CANADA

Rob Feltham (ETPG)
NFER-Nelson, ENGLAND

Rocio Fernandez-Ballesteros (EAPA)
Universidad de Autonoma, SPAIN

John H. A. L. de Jong (ILTA)
CITO, THE NETHERLANDS

Ingrid Munck (IEA)
Statistics Sweden, SWEDEN

José Muñiz (ITC)
Universidad de Oviedo, SPAIN

Ype Poortinga (IACCP)
Tilburg University, THE NETHERLANDS

Isik Savasir (IUPsyS)
Hacettepe University, TURKEY

Charles Spielberger (IAAP)
University of South Florida, U.S.A.

Fons van de Vijver (ITC)
Tilburg University, THE NETHERLANDS

Jac N. Zaal (ITC)
GITP International, THE NETHERLANDS

Research Associate

Anil Kanjee (ITC)
University of Massachusetts at Amherst, U.S.A.

Figure 2. Committee members and the organizations they represent.

Context

C.1 Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.

C.2 The amount of overlap in the constructs in the populations of interest should be assessed.

Instrument Development and Adaptation

D.1 Instrument developers/publishers should insure that the adaptation process takes full account of linguistic and cultural differences among the populations for whom adapted versions of the instrument are intended.

D.2 Instrument developers/publishers should provide evidence that the language use in the directions, rubrics, and items themselves as well as in the handbook are appropriate for all cultural and language populations for whom the instrument is intended.

D.3 Instrument developers/publishers should provide evidence that the choice of testing techniques, item formats, test conventions, and procedures are familiar to all intended populations.

D.4 Instrument developers/publishers should provide evidence that item content and stimulus materials are familiar to all intended populations.

D.5 Instrument developers/publishers should implement systematic judgmental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and compile evidence on the equivalence of all language versions.

D.6 Instrument developers/publishers should ensure that the data collection design permits the use of appropriate statistical techniques to establish item equivalence between the different language versions of the instrument.

D.7 Instrument developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the instrument, and (2) identify problematic components or aspects of the instrument which may be inadequate to one or more of the intended populations.

Figure 3. Draft of the instrument adaptation guidelines.

Figure 3, continued:

D.8 Instrument developers/publishers should provide information on the evaluation of validity in all target populations for whom the adapted versions are intended.

D.9 Instrument developers/publishers should provide statistical evidence of the equivalence of questions for all intended populations.

D.10 Non-equivalent questions between versions intended for different populations should not be used in preparing a common scale or in comparing these populations. However, they may be useful in enhancing content validity of scores reported for each population separately.

Administration

A.1 Instrument developers and administrators should try to anticipate the types of problems that can be expected, and take appropriate actions to remedy these problems through the preparation of appropriate materials and instructions.

A.2 Instrument administrators should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores.

A.3 Those aspects of the environment that influence the administration of an instrument should be made as similar as possible across populations for whom the instrument is intended.

A.4 Instrument administration instructions should be in the source and target languages to minimize the influence of unwanted sources of variation across populations.

A.5 The instrument manual should specify all aspects of the instrument and its administration that require scrutiny in the application of the instrument in a new cultural context.

A.6 The administrator should be unobtrusive and the administrator-examinee interaction should be minimized. Explicit rules that are described in the manual for the instrument should be followed.

Figure 3, continued:

Documentation/Score Interpretations

I.1 When an instrument is adapted for use in another population, documentation of the changes should be provided, along with evidence of the equivalence.

I.2 Score differences among samples of populations administered the instrument should not be taken at face value. The researcher has the responsibility to substantiate the differences with other empirical evidence.

I.3 Comparisons across populations can only be made at the level of invariance that has been established for the scale on which scores are reported.

I.4 The instrument developer should provide specific information on the ways in which the socio-cultural and ecological contexts of the populations might affect performance on the instrument, and should suggest procedures to account for these effects in the interpretation of results.

Guideline D.1: General and Professional Requirements

Instrument developers/publishers should insure that the adaptation process takes full account of linguistic and cultural differences among the populations for whom adapted versions of the instrument are intended.

Rationale/Explanation

The expertise and experience of translators is perhaps the most crucial aspect of the entire process of adapting instruments as it can significantly affect the reliability and validity of the instrument (Bracken & Barona, 1991). For example, translators without domain specific or technical knowledge often resort to literal translations that may cause misunderstanding in the target population and threaten the validity of the instrument (Hambleton & Kanjee, in press). Consequently, the selection of appropriately qualified translators is an important aspect of the instrument adaptation process. While expertise in both languages is a basic condition, familiarity and experience with (1) both cultures, (2) the content of the "subject area," and (3) the principles of developing instruments, should also be included as part of the essential requirements for the selection and/or training of translators. A single individual can hardly be expected to have all of the required qualities, therefore in general it is recommended that a team of specialists be formed to accomplish an accurate adaptation.

Steps to Meet the Guideline

1. As a basic minimum, ensure that translators are qualified and experienced in the source and target languages as well as in both cultures (Butcher & Garcia, 1978). Certification and/or prior experience is an important requirement. For instance, it cannot be assumed that bilinguals have equal command of both languages in all relevant domains or are equally familiar with both cultures.

2. Knowledge of the subject matter is an important requirement for any translator involved in adapting a testing instrument. Without at least some content knowledge, the subtleties and nuances of the subject matter can be lost. Prior familiarization with the subject matter for translators lacking domain specific knowledge should be included as part of the instrument adaptation process.

Figure 4. An example of guideline D.1 in its complete form (i.e., with all four sections included).

Figure 4 continued:

3. The translators selected should possess some basic knowledge about instrument development and item writing (Bracken & Barona, 1991). Test translators need to know, for example, that options in a multiple choice question should be of comparable length, that associations that might lead test-wise examinees to the correct answer should be avoided, and that unobservant translation of distractors in multiple choice items may lead to two or more distractors having the same meaning in the target language (Hambleton & Kanjee, in press). In the case where translators lack this expertise, a training session should be included in the instrument adaptation process. One example of the type of problem that might arise (brought to our attention by Anita Wester from the University of Umea in Sweden) is this question:

   Where is a bird with webbed feet most likely to live?

   a. *in the mountains*
   b. *in the woods*
   c. *in the sea*
   d. *in the desert*

   When this question was translated into Swedish, "webbed feet" became "swimming feet," which then provided an obvious clue to Swedish children about the location of the correct answer.

4. An adaptation project should be carried out by a team of specialists. Translators should participate in such a project team and be involved in the decision making process, and their opinions and views should be actively sought and acknowledged. According to Brislin (1986), this approach can greatly improve the quality of an adaptation. The teamwork approach can help to (1) enable the use of the back-translation methods (see step 5, below); (2) allow translators to compare and discuss their work and thus improve on the relevance and quality of translations; and (3) can help to ensure that specialist knowledge in all required fields is accessible.

5. One possible procedure is to use a team of bilingual translators working independently or in small groups to adapt the instrument. Another procedure is the use of monolingual instrument developers and bilingual translators simultaneously, where instruments are first translated/adapted by a bilingual, rewritten by a monolingual instrument developer and then re-assessed by a bilingual (Brislin, 1986). Brislin (1986) notes that the advantage of this procedure is that "monolingual

Figure 4 continued:

translators" can rewrite instruments so that it would be clear to native speakers, and it controls for situations where the target version is assumed to be good, even though it is possible for highly skilled translators to produce a better back-translated version than the mangled target version. In the case where only a single translator is available, the use of a member from the target language population to assist the translator is recommended. Thus the translator can at least discuss the product with members from the source as well as target languages.

Common Errors

1. Selection of translators or easily available individuals familiar to the instrument developer (i.e. friends), simply because they are bilingual has been shown to be an unsuccessful practice (Brislin, 1970).

2. Failure to ensure that translators selected are familiar with the content area as well as experienced in instrument development.

References for Additional Study

Bracken, B. A, & Barona, A. (1991). State of the art procedures for translating, validating and using psycho-educational tests in cross-cultural assessment. School Psychology International, 12, 119-132.

Brislin, R. (1970). Back-translation for cross-cultural research. Journal of Cross-Cultural Psychology, 1, 185-216.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds), Field methods in cross-cultural psychology (pp 137-164). Newbury Park, CA: Sage Publishers.

Butcher, J. N., & Garcia, R. E. (1978). Cross-national application of psychological tests. The Personnel and Guidance Journal, 56(8), 472-475.

Hambleton, R. K., & Kanjee, A. (In press). Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods. In T. Husen & T. N. Postlewaite (Eds), International Encyclopedia of Education (2nd ed.). Oxford, UK: Pergamon Press.

Prieto, A. J. (1992). A method for translation of instruments to other languages. Adult Education Quarterly, 43, 1-14.